Data Collection and Preprocessing Phase

Date	12 June 2025
Team ID	SWTID1749627644
Project Title	Human Resource Management: Predicting Employee Promotions using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing:

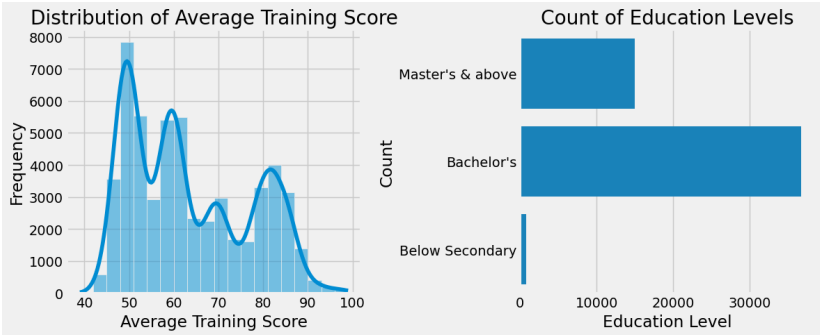
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

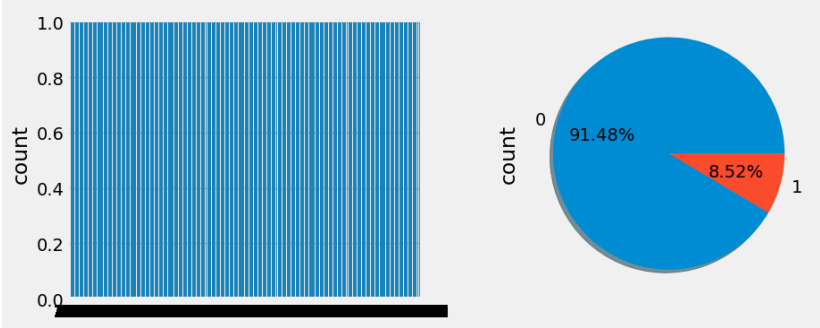
Section	Description
Data Overview	<u>Dimension:</u> 54808 rows x 14 columns <u>Descriptive Statistics:</u>

	employee_id	no_of_trainings	age	previous_year_rating	length_of_service
count	54808.000000	54808.000000	54808.000000	50684.000000	54808.000000
mean	39195.830627	1.253011	34.803915	3.329256	5.865512
std	22586.581449	0.609264	7.660169	1.259993	4.265094
min	1.000000	1.000000	20.000000	1.000000	1.000000
25%	19669.750000	1.000000	29.000000	3.000000	3.000000
50%	39225.500000	1.000000	33.000000	3.000000	5.000000
75%	58730.500000	1.000000	39.000000	4.000000	7.000000
max	78298.000000	10.000000	60.000000	5.000000	37.000000

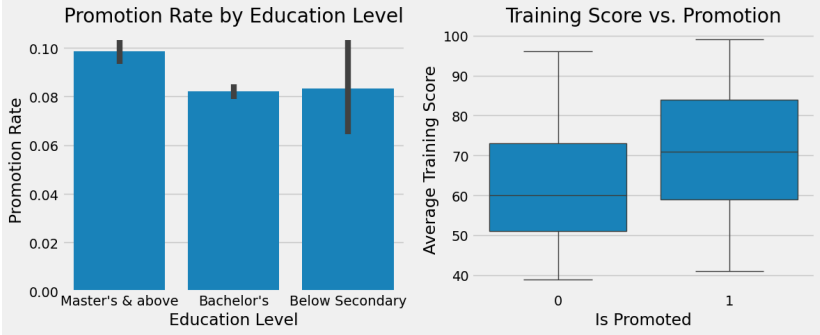
KPIs_met >80%	awards_won?	avg_training_score	is_promoted
54808.000000	54808.000000	54808.000000	54808.000000
0.351974	0.023172	63.386750	0.085170
0.477590	0.150450	13.371559	0.279137
0.000000	0.000000	39.000000	0.000000
0.000000	0.000000	51.000000	0.000000
0.000000	0.000000	60.000000	0.000000
1.000000	0.000000	76.000000	0.000000
1.000000	1.000000	99.000000	1.000000

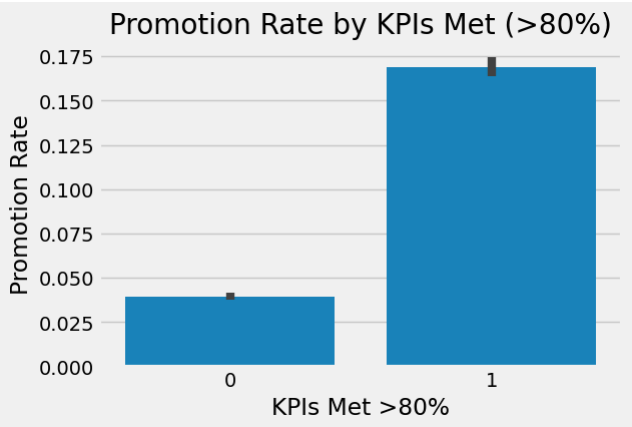
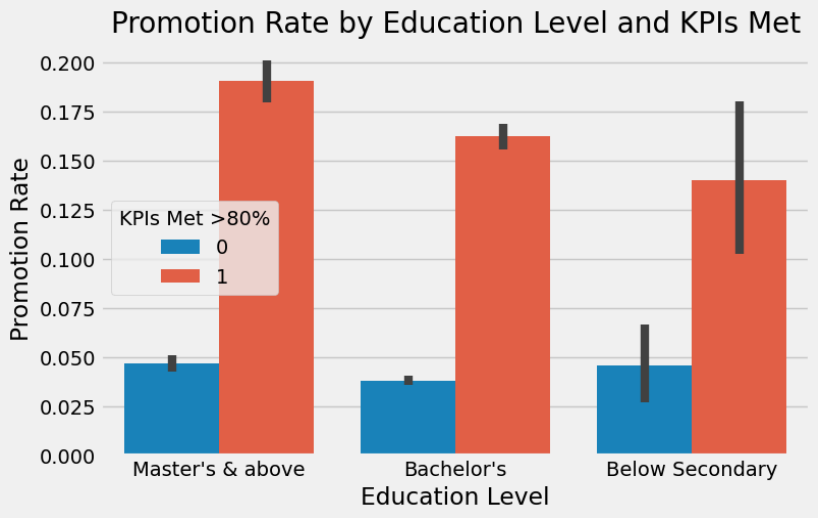
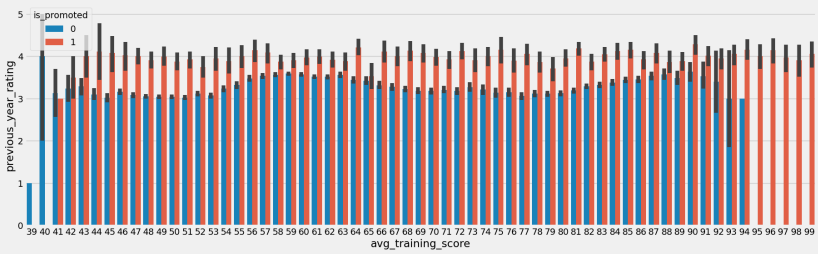
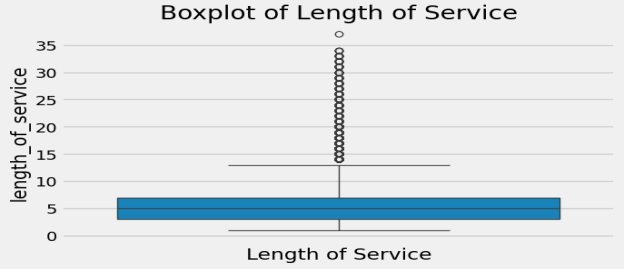
Univariate Analysis





Bivariate Analysis



	
Multivariate Analysis	 
Outliers and Anomalies	

```
Number of outliers: 3489
```

	length_of_service	is_promoted
13	16	0
42	26	0
60	17	1
74	14	0
99	17	0
...	...	...
54691	19	0
54695	18	1
54697	15	0
54754	14	0
54803	17	0

3489 rows × 2 columns

Data Preprocessing Code Screenshots

Loading Data

```
df = pd.read_csv('../Dataset/emp_promotion.csv')
print('Shape of train data {}'.format(df.shape))
```

✓ 0.1s

```
Shape of train data (54808, 14)
```

```
df.head()
```

✓ 0.0s

	employee_id	department	region	education	gender	recruitment_channel
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing
1	65141	Operations	region_22	Bachelor's	m	other
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing
3	2542	Sales & Marketing	region_23	Bachelor's	m	other
4	48945	Technology	region_26	Bachelor's	m	other

Handling Missing Data	```python df.isnull().sum()```
	``
	department 0
	education 2409
	no_of_trainings 0
	age 0
	previous_year_rating 4124
	length_of_service 0
	KPIs_met >80% 0
	awards_won? 0
	avg_training_score 0
	is_promoted 0
	dtype: int64
	``
	```python print(df['education'].value_counts()) df['education'] = df['education'].fillna(df['education'].mode()[0])```
	``
	education
	Bachelor's 36669
	Master's & above 14925
	Below Secondary 805
	Name: count, dtype: int64
	``
	```python print(df['previous_year_rating'].value_counts()) df['previous_year_rating'] = df['previous_year_rating'].fillna(df['previous_year_rating'].mode()[0])```
	``
	previous_year_rating
	3.0 18618
	5.0 11741
	4.0 9877
	1.0 6223
	2.0 4225
	Name: count, dtype: int64
	``
	Python

Data Transformation	We do not need employee id, gender, region, recruitment channel attributes for predicting promotion, so we will be removing these unwanted features.
	```python df = df.drop(['employee_id','gender','region','recruitment_channel'], axis=1)```
	Python
	Removing inconsistent rows
	```python df.drop(index=[31860,51374], inplace=True)```
	Capping outliers
	```python df['length_of_service']=[upperBound if x > upperBound else x for x in df['length_of_service']]```
	Python

Feature Engineering	```python # Feature mapping on education column df['education'] = df['education'].replace(("Below Secondary", "Bachelor's", "Master's & above"), (1,2,3)``` Python

	``python
	lb = LabelEncoder()
	df['department'] = lb.fit_transform(df['department'])
	``
	``python
	sm = SMOTE()
	x_resample, y_resample = sm.fit_resample(x,y)
	``

Save Processed Data	-